De quelles façons l'intelligence artificielle se sert-elle des neurosciences ?

Frédéric Alexandre, Université de Bordeaux, Inria-Institut des Maladies Neurodégénératives

L'Intelligence Artificielle (IA) s'est construite sur une opposition entre connaissances et données. Les neurosciences ont fourni des éléments confortant cette vision mais ont aussi révélé que des propriétés importantes de notre cognition reposent sur des interdépendances fortes entre ces deux concepts. Cependant l'IA reste bloquée sur ses conceptions initiales et ne pourra plus participer à cette dynamique vertueuse tant qu'elle n'aura pas intégré cette vision différenciée.

IA symbolique et numérique

La quête pour l'IA s'est toujours faite sur la base d'une polarité entre deux approches exclusives, symbolique ou numérique. Cette polarité fut déclarée dès ses origines, avec certains de ses pères fondateurs comme J. von Neumann ou N. Wiener proposant de modéliser le cerveau et le calcul des neurones pour émuler une intelligence, et d'autres comme H. Newell ou J. McCarthy soulignant que, tout comme notre esprit, les ordinateurs manipulent des symboles et peuvent donc construire des représentations du monde et les manipulations caractérisques de l'intelligence. Cette dualité est illustrée par l'expression des frères Dreyfus « Making a Mind versus Modelling the Brain », dans un article (Dreyfus & Dreyfus, 1991) où ils expliquent que, par leur construction même, ces deux paradigmes de l'intelligence sont faits pour s'opposer : Le paradigme symbolique met l'accent sur la résolution de problèmes et utilise la logique en suivant une approche réductionniste et le paradigme numérique se focalise sur l'apprentissage et utilise les statistiques selon une approche holistique.

On connaît la suite de l'histoire avec, tour à tour, chaque approche écrasant l'autre à l'occasion du succès éclatant d'une technique particulière, suivi de désillusions entraînant ce que l'on appelle un hiver de l'IA. Aujourd'hui, l'IA a fait des progrès indéniables, mais nous subissons toujours cette dualité, même si le vocabulaire a un peu évolué et que l'on parle maintenant d'IA basée sur les connaissances (pour le web sémantique) ou sur les données (et les *data sciences*). Nous sommes actuellement sans conteste dans une période numérique où tout le monde n'a que le Deep Learning à la bouche, même si des voix commencent à s'élever pour prédire une chute proche si l'on n'est pas capable d'associer ces techniques numériques à une interprétabilité (Lipton, 2017), permettant transparence et explications, deux notions du monde des connaissances.

Sommes-nous encore partis pour un cycle, à toujours nous demander laquelle de ces deux approches finira par démontrer qu'elle était la bonne solution, ou saurons-nous sortir du cadre et trancher le nœud gordien ? C'est dans cette dernière perspective que je propose de revenir aux fondamentaux. Puisque les deux approches s'accordent au moins sur le fait qu'elles cherchent à reproduire nos fonctions cognitives supérieures, ne devrait-on pas commencer par se demander si notre cognition est symbolique ou numérique ?

Mémoires implicite et explicite dans le cerveau

A cette question, les Sciences Cognitives répondent d'abord « les deux » et soulignent (Squire, 2004) que notre mémoire à long terme est soit explicite soit implicite. D'une part nous pouvons nous souvenir de notre repas d'hier soir (mémoire épisodique) ou avoir la connaissance que le ciel est bleu (mémoire sémantique) ; d'autre part nous avons appris notre langue maternelle et

nous pouvons apprendre à faire du vélo (mémoire procédurale). Nous savons que (et nous en sommes conscients, nous savons l'expliquer) ou nous savons faire (et nous pouvons en faire la démonstration, sans être capable de ramener cette connaissance au niveau conscient). On retrouve ici les principes décrits respectivement en IA par la manipulation explicite de connaissances ou implicite de données.

Les neurosciences ont identifié des circuits cérébraux correspondants, avec en particulier les boucles entre les ganglions de la base et le cortex plutôt impliquées dans la mémoire implicite, et l'hippocampe et ses relations avec l'ensemble du lobe temporal médial, essentiel pour la mémoire explicite. Les deux modes d'apprentissage sont à l'œuvre dans deux phénomènes : La consolidation et la formation des habitudes.

Les mécanismes de la consolidation

Ces mémoires complémentaires sont construites avec un apprentissage lent et procédural dans le cortex et la formation rapide d'associations arbitraires dans l'hippocampe (McClelland et al., 1995). Prenons un exemple : allant toujours faire mes achats dans le même supermarché, je vais former, après de nombreuses visites, une représentation de son parking, mais à chaque visite, je dois aussi me souvenir de l'endroit précis où j'ai laissé ma voiture. Les modèles computationnels permettent de mieux comprendre ce qui est à l'œuvre ici. Les modèles d'apprentissage procédural implicite, généralement en couches, montrent que des régularités sont extraites statistiquement, à partir de nombreux exemples dont les représentations doivent se recouvrir pour pouvoir généraliser. Mais si l'on souhaite apprendre ensuite des données avec d'autres régularités, on va observer l'oubli catastrophique des premières relations apprises.

Inversement, dans un modèle d'apprentissage explicite de cas particuliers, généralement avec des réseaux récurrents, on va privilégier le codage de ce qui est spécifique plutôt que de ce qui est régulier dans l'information (pour retrouver ma voiture, je ne dois pas généraliser sur plusieurs exemples mais me souvenir du cas précis). Cet apprentissage sera plus rapide, puisqu'on ne cherchera pas à se confronter à d'autres exemples mais à apprendre par cœur un cas particulier. Mais l'expérimentation avec ce type de modèles montre des risques d'interférence si on apprend trop d'exemples proches, ainsi qu'un coût élevé pour le stockage des informations (ce qui n'est pas le cas pour l'apprentissage implicite). Il est donc impératif de limiter le nombre d'exemples stockés dans l'hippocampe.

Des transferts de l'hippocampe vers le cortex (que l'on appelle consolidation, se produisant principalement lors des phases de sommeil) traitent les deux problèmes évoqués plus haut. D'une part, lorsque des cas particuliers proches sont stockés dans l'hippocampe, leurs points communs sont extraits et transférés dans le cortex. D'autre part, l'hippocampe, en renvoyant vers le cortex des cas particuliers, lui permet de s'entrainer de façon progressive, en alternant cas anciens et nouveaux et lui évite l'oubli catastrophique.

Les mécanismes de la formation des habitudes

La prise de décision peut se faire selon deux modes, réflexif et réflectif (Dolan & Dayan, 2013), tel que proposé historiquement par les behavioristes pour qui le comportement émergeait implicitement d'un ensemble d'associations Stimulus-Réponse et par les cognitivistes qui imaginaient plutôt la construction de cartes cognitives où des représentations intermédiaires explicites étaient exploitées. Là aussi, les apprentissages implicite et explicite sont à l'œuvre. Pour prendre une décision, une représentation explicite du monde permettra de façon

prospective d'anticiper les conséquences que pourraient avoir nos actions et de choisir la plus intéressante. Avec sa capacité à former rapidement des associations arbitraires, l'hippocampe semble massivement impliqué dans la construction de ces cartes cognitives explicites.

Ensuite, après avoir longuement utilisé cette approche dirigée par les buts, on peut se rendre compte, par une analyse rétrospective portant sur de nombreux cas, que dans telle situation la même action est toujours sélectionnée, et se former une association situation-action dans le cortex par apprentissage lent, sans se représenter explicitement le but qui motive ce choix. On appelle cela la formation des habitudes.

Mais que fait l'IA ?

La dualité implicite/explicite a conforté l'IA dans ses aspects numériques/symboliques ou basés sur les données et sur les connaissances. L'IA n'a cependant pas intégré un ensemble de résultats qui montrent que, au delà d'une simple dualité, les mémoires implicites et explicites interagissent subtilement pour former notre cognition.

Concernant la consolidation, l'hippocampe est en fait alimenté presque exclusivement par des représentations provenant du cortex, donc correspondant à l'état courant de la mémoire implicite, ce qui indique que ces deux mémoires sont interdépendantes et co-construites. Comment ces échanges se réalisent entre le cortex et l'hippocampe et comment ils évoluent mutuellement restent des mécanismes très peu décrits et très peu connus en neurosciences.

Concernant la formation des habitudes, cette automatisation de notre comportement n'est pas à sens unique et nous savons figer un comportement puis le réviser par une remise en cause explicite quand il n'est plus efficace puis le reprendre si besoin. Là aussi, ces mécanismes sont très peu compris en neurosciences.

La modélisation a été une source d'inspiration pour aider les neurosciences à formaliser et à décrire les mécanismes de traitement de l'information à l'œuvre dans notre cerveau. Pourtant, concernant ces modalités d'associations flexibles entre nos mémoires implicites et explicites, l'IA ne joue pas son rôle d'aiguillon pour aider les neurosciences à avancer sur ces questions, car elle reste bloquée sur cette dualité rigide et stérile entre données et connaissances, alors que les relations entre connaissances et données devraient être au cœur des préoccupations d'une IA soucieuse de résoudre ses points de blocage. Il est donc temps d'exposer au grand jour ce hiatus et de demander à l'IA de jouer son rôle d'inspiration.

Dreyfus H.L., Dreyfus S.E. (1991) Making a Mind Versus Modelling the Brain: Artificial Intelligence Back at the Branchpoint. In: Negrotti M. (eds) *Understanding the Artificial: On the Future Shape of Artificial Intelligence. Artificial Intelligence and Society*. Springer, London.
https://link.springer.com/chapter/10.1007/978-1-4471-1776-6_3

Lipton, Z. C. (2017). *The Mythos of Model Interpretability*. http://arxiv.org/abs/1606.03490

Squire, L. R. (2004). Memory systems of the brain : a brief history and current perspective. *Neurobiology of Learning and Memory*, 82, 171–177.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457.

Dolan, R. J., & Dayan, P. (2013). Goals and Habits in the Brain. *Neuron*, *80*(2), 312–325. https://doi.org/10.1016/j.neuron.2013.09.007